A Scalable Recirculating Shuffle Network with Deflection Routing

S. P. Monacos
High Speed optical Systems Group
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

and

A. A. Sawchuk
Signal and Image Processing Institute
Department of Electrical Engineering
University of Southern California
Los Angeles, CA. 90089-2564

Submitted to IEEE Journal on Selected Areas in Communications, March 29, 1995

Abstract

We describe a high-speed optical wide area network (WAN) architecture called the multi-cylinder ShuffleNet (MCSN). Its basic topology is a recirculating snuffle exchange network with an all-optical data path from source to destination. The network capacity and performance is improved by augmenting it with multiple parallel copies of the original topology called routing cylinders. Packets are dynamically stored within the switching nodes and links of the network, and congestion control within the network is handled by routing blocked packets onto alternative, routing cylinders. The architecture is easily scalable and is designed for bit-synchronous, packet-asynchronous traffic, meaning that incoming packets need only be reclocked for bit alignment but not packet alignment. The network utilizes a simple distributed control scheme which matches the rate at which packets can be input into the network to the rate at which packets can be transmitted through the network (without congestion), and may be built using current electronic and/or optical/electrooptic technology. Simulation results which demonstrate these features are given.

1 Introduction

Optical interconnection networks offer the potential for bandwidth in the tens to hundreds of gigabits/sec over a given data path. Their primary limitation is that complex logic, data buffering and routing is difficult to perform in the optical domain. Our goal is to perform these functions as fast as possible while maintaining the data in all-optical form as much as possible to avoid electronic/optical conversions.

Current optical communication networks fall into two basic categories: i) a ring or bus topology with fiber interconnects; and ii) circuit switched networks with optical point-to-point data paths and electronic paths within switching nodes, A ring or bus topology is simple to implement and maintains data in optical form while in the network, but has only one data path for all communications. Circuit switched networks have many paths, but the data rate on each is limited by optoelectronic conversions.

As computational and communications demands increase. such networks are inadequate to handle future traffic loads. In order to meet ever-increasing bandwidth requirements, we describe a network architecture that is suitable for optical implementation because routing decisions arc simple local operations, and buffering is done by fiber loops and fiber inter–node connections, The network is composed of an interconnected set of *switching nodes* arranged in a *Multi–Cylinder ShuffleNet* (MCSN) topology that addresses many of the implementation issues while providing good performance characteristics through scalability of the network. The emphasis here is to keep the control mechanisms simple by designing scalability into every aspect of the network architecture. We begin by defining some relevant terms used in this work.

1.1 Terminology

A switching *node* of the network consists of J physical input ports and J physical output ports and is used to route data through the network based on the packet header. A *host* refers to a device which can transmit and/or receive packets from a switching node of the network. The *input bandwidth* of a switching node is the maximum rate that packets can be transmitted from a host to its associated switching node. The *output bandwidth* is the maximum rate that packets can be sent from a switching node to its associated host.

We define a *recirculating network* as a network topology in which any switching node of the network can transmit packets to any other switching node of the network (although some packets may travel through intermediate nodes to reach their destination). We further define a *superset network* architecture as multiple copies of a recirculating network topology. Additionally, these copies share a common set of nodes to allow packets to route between the copies.

A *packet asynchronous* system does not require that packet headers (and the packets themselves) be synchronized in time. While the architectures described here are packet asynchronous, they do require *bit synchronous* data, i.e., *bits* flowing through the network are clocked synchronously through the switching nodes.

A network *protocol defines the* rules in routing a packet through a network. A *store-and-forward* protocol avoids misrouting a header or packet by storing the header or packet in memory until the desired path is available. A *defection* protocol never stores a header or packet and results in a misroute or deflection of the header or packet through another path if the desired path is not available.

A network is called *circuit switched* if a path from a source to a destination must be established before any data is sent. A network is *packet switched* if no path setup is needed and routing is based only on the packet header and local traffic conditions.

1.2 Signal Format

We define a data *packet* as an indivisible unit of information which may exist as a serial time signal or wavelength-parallel signal as shown in Fig. 1. A serial packet shown at the top of Fig. 1 consists of an n–bit binary packet header; a d-bit payload or data for this header; and a *t*-bit trailer marking the end of the packet. The header contains the address of the desired destination host. A *message* contains the total communication from a source host to a receiving host and is defined as a sequence of packets.

The serial packet format is suitable for packet switched networks which can handle long duration packets. These networks work best when there are buffers in the switching nodes for storing packets that are blocked due to output port contention [1],[2]. For all-optical data path networks, such buffering capability is not currently available at high data rates.

In order to reduce the probability of packet deflections within the network, it is desirable to reduce the packet duration by parallelizing the header, data and trailer information as shown in Fig. 1. For an all–optical data path network, this scheme is realized by using a wave.length-parallel or wavelength division multiplex (WDM) format in which the n header bits arc encoded at λ_H, the d data bits arc encoded at λ_1 through λ_M, and the clock is encoded at λ_C. The duration of one bit is defined as a time slot, and three clock pulses (time slots) which would correspond to a three bit header arc shown as an example at the bottom of Fig. 1. The start and stop bits mark the boundaries of the packet and arc encoded as double intensity clock pulses at λ_C. In this WDM implementation, the trailer bits arc not used, and their role is served by the start/stop bits. This packet format simplifies the maintenance of internal switch states within a switching node for the duration of the packet.

The packets shown in Fig. 1 are rapidly routed on the fly based on the packet header. The amount of data which can be transferred in a single packet depends on the size of a data word and the total number of data words (i.e. wavelengths) used for transmitting data. If a message cannot be encapsu-

lated into one packet, the source host generates enough packets with similar headers to transmit the entire contents of the message. These packets are sent in a time sequential fashion until the last packet is transmitted to the network.

The remainder of this paper is organized as follows. Section 2 describes the desired optical network protocol. Section 3 reviews various network architectures suitable to this protocol. Section 4 discusses the multi-cylinder SN and its internal routing algorithm. Section 5 presents simulation results for the single and multi-cylinder SN configurations. Conclusions are given in section 6.

2 Optical WAN Protocol

An optical WAN presents several constraints to the system designer due to the lack of static optical storage and the geographic extent of the network. These constraints are: i) packets from geographically distributed sources cannot be synchronized; ii) retiming of optical packets is difficult; iii) packets must be stored dynamically within the pathways of the network; iv) packets must be large enough to encapsulate several bytes (i.e. greater than a few bits); and v) the flow control mechanism cannot use a global request/acknowledge because the latency of a WAN with gigabit/see data rates is on the order of milliseconds. The short packet protocol and network architecture described here specifically addresses these issues and accommodates deflection routing and load imbalances without the use of a global control mechanism. The remainder of this section discusses the desired network protocol and architectural issues for the proposed network.

2.1 General Network Protocol Issues

Constraints i) through iii) listed above result in *asynchronous* packet traffic flowing through the network. If a packet requires a data path currently in use, the packet cannot be stored until the path is available but instead must be routed through an alternative path. Thus, our deflection protocol must be capable of handling asynchronous packet traffic. Constraints iv) and v) are a result of the lack

of buffering in an optical WAN, Instead of u sing input and output buffers in the network switching nodes, we usc short duration packets to reduce the blocking probability y within the network [1],[2].

The MCSN network protocol combines several advantages to maximize WAN performance with reduced physical complexity. It u ses deflection routing to avoid buffering and simplify the hardware, a wavelength division multiplexed (WDM) packet for mat to utilize fully the optical link bandwidth, and asynchronous packets like that used in asynchronous transfer mode (ATM) networks. Thus, these networks route on the fly using the packet header, with no path setup or tear down, to avoid the need for a global request/acknowledge mechanism.

However, the protocol requires that traffic pattern imbalances for a sustained period must be controlled by some form of *source throttling* at the inputs to the network [3], [4] and *bandwidth augmentation* at the outputs of the network. The former implies regulation of the rate at which various sources can input traffic into the network based on expected traffic statistics. The latter means that allowable bandwidth from a switching node to its associated host must be greater than the bandwidth from a host to its associated switching node, This augmentation is realized by using additional ports to transmit packets from a node to its host.

3 Recirculating Architectures

In this section wc discuss two basic optical packet network topologies as an introduction to the MCSN concept. Rcfcrcncc [9] has additional background on related networks.

3.1 Manhattan Street Network

The Manhattan Street Network (MSN) is a grid of crossbar nodes connected in a mesh topology with unidirectional links. Adjacent rows and columns are oriented so that data flows in opposite directions similar to the. one-way streets in Manhattan [1]. The wp around links at the edges forma log-

ical torus [1]. In this recirculating architecture, each node is a 3x3 crossbar with a pair of 1/0 ports used to connect a host to each node as shown in Fig. 2.

The MSN is highly connected in the sense that there are multiple paths having the same number of links between nodes, and this property may reduce the number of packet deflections by providing alternative paths without increasing the number of links to be traversed [1]. In addition, the penalty for a deflection is never more than four links since a packet can go "around the block" [1]. Also, the number of turns required to reach a destination is never more than three [1]. As the network grows in size, the probability that a packet must tum decreases, which also decreases the probability of deflections [1]. Finally, if a packet is deflected, it can use either output link at the next node to "go around the block" [1].

Rather than trying to minimize the extent of the network, this architecture maintains a high degree of regularity in the topology. This approach reduces the probability of' a deflection at the expense of increasing the shortest paths in the network. [1], increasing the path lengths increases the link uti-lization and reduces the maximum network throughput [1]. Hence the generic MSN is a compromise between reducing the effect of deflections and network throughput.

3.2 ShuffleNet

The ShuffleNet (SN) [5], [6] is based on the omega or multistage shuffle exchange network. The omega network is unidirectional and consists of a cascade of k stages of **2x2** crossbar switching nodes. Figure 3 shows an example of a 3–stage (k=3) omega network. Data from 2^k host devices enters from the left side of the network and undergoes a shuffle operation between switch stages [7],[8]. Each stage of a k-stage omega network has 2^{k-1} 2x2 switching nodes for a total of $k2^{k-1}$ switching nodes with 2^k input lines incident at the first stage ports. After k switch stages, data emerges from the 2^k output ports.

The SN is defined by using k stages of 2^k switching nodes per stage, with a perfect shuffle interconnection between stages, and connecting the outputs of this modified omega network to the inputs to realize a recirculating topology. In addition, each 2x2 crossbar node is replaced by a 3x3 crossbar node, so that the added input and output ports at each node are connected to a separate host (processor) as shown in Fig, 4. Fork stages of switching nodes, the total number of switching nodes for the SN topology is $k2^k$ [5], [6]. An example is shown in Fig. 5 for an 8-node (k=2) network [1],[5],[6].

This topology minimizes the maximum path length of any two-connected network [1], [5]. That is, the maximum number of hops needed to route a packet from a given source node to an y destination node is minimized, where a hop corresponds to passage of the packet through an intermediate node while routing to the destination node. The penalty paid for this characteristic is that a deflection due to contention can result in an increase of k extra hops for a k-stage SN [6]. Consequently, efficient utilization of the SN topology requires avoidance of packet deflections, Augmentation of the SN topology to avoid packet deflections is the main focus of this work.

4 Topological Considerations

Wc first define several terms that arc useful in discussing the effects of net work topology on network throughput. The term *capacity* denotes the number of bits of information that can be transferred per unit time through a link or group of links. The *input capacity* of a network is defined as the sum of the input data rates from all hosts on all ports to their respective switching, nodes in the network. The *internal network capacity is the* aggregate data rate of all links within the network without regard to possible packet deflections and output port contention,

The MCSN architecture was developed with scalability being the most important design requirement. One way of increasing the internal network capacity with respect to the input capacity is to increase the number of SN switching nodes while keeping the number of host connections fixed.

Figure 5 shows an 8–node SN with all eight switching nodes connected to host devices. Here, we could add one or more additional stages of switching nodes that have no host devices connected.

While this approach may alleviate some congestion problems, it does not *scale the* basic SN topology of section 3.2 in an optimal way to increase network capacity. Because deflected packets in a *k* stage SN may incur an additional *k* hops in reaching their destination, increasing the number of stages of switching nodes (and therefore *k)* results in a larger penalty for a packet deflection. Thus, adding network stages in this way helps only if packet deflection is avoided. If network "hot spots" exist and result in packet deflections, the end result is to increase the latency of deflected packets due to the additional stages. In contrast, the MSN topology is better suited to this form of scaling due to the minimal penalty associated with packet deflections. However, this characteristic is realized at the expense of increasing the average path length (as defined by the expected number of hops) between nodes [1].

4.1 **Multi-Cylinder ShuffleNet**

Another way to improve performance is to parallelize the interconnections between the nodes around the cylinder of a SN to rnakc a multi-cylinder ShuffleNet (MCSN). A *routing cylinder is* defined as a set of node-to-node links that provides the perfect shuffle interconnections between stages of nodes in the SN. An R-cylinder MCSN has *R* parallel perfect shuffle interconnections between stages of nodes in the SN, rather than only one interconnection. Thus an 8–node, R-cylinder SN is topologically equivalent to the single-cylinder SN in Fig. 5, but each node--to-node link is expanded to *R physical data ports*. Figure 6 shows a general MCSN node. Here wc expand the number of ports at each node as shown in Fig. 4 so that there are a total of 2R network input ports and 2R network output ports as shown in Fig. 6. In addition, there maybe *Q* input and output ports for internal recir-

culating links within anode, and a total of H output ports to an associated host for increased bandwidth. We keep the number of host input ports at one.

This MCSN node internally uses a *Permutation Engine* (PE), which is a scalable strictly non-blocking crossbar mesh topology containing a simple distributed internal routing control mechanism. PEs are described in detail in Ref. [9], PEs perform non-blocking routing with bit-synchronous but packet-asynchronous data (incoming packets must be bit-aligned but need not be packet-aligned). For the R-cylinder SN, the PE switching nodes also provide dynamic routing of packets between SN cylinders. This approach alleviates packet congestion associated with a single-cylinder SN by routing blocked packets onto alternative routing cylinders to avoid packet deflections.

Thus, the total number of input data ports for the MCSN node in Fig. 6 is $2R+Q+1$, and the total number of output data ports is $2R+Q+H$. The PE is designed with an equal number of inputs and outputs, thus $H-1$ unused input ports are left free. By using H data paths from the node to its host, we allow for up to H other hosts simultaneously to send packets to this host on a steady-state basis. By connecting Q output data paths to Q input data paths, packets which are initially blocked from the host at the destination node can try again without recirculating through the whole MCSN. The emphasis of this design approach is to avoid packet deflections at the expense of increasing the mean routing delay for the augmented MCSN. As we show in section 4.2, this extra delay does not decrease the network capacity.

4.2 **Network Analysis**

To analyze the performance of this network, we define several variables: R is the number of parallel SN cylinders, and N is the total number of SN switching nodes. We also define τ_H as the header detection/decoding delay time, τ_D as the node routing delay time, and τ_L as the internode link delay time. For simplicity, all delays are measured in units of *simulation time slots*, corresponding to the dura-

tion of one bit time in a packet as shown in Fig. 1. Packets maybe stored within the switching node fabric and in the data paths between nodes, thus a total of $RN(\tau_H+\tau_D+\tau_L)$ time slots are available to store packets in the network in a non-blocking fashion.

Finally, we define τ_P as the packet length, measured in units of simulation time slots per packet. We allow for one additional time slot for the dead time between packets. This sum τ_P+1, is the average number of slots between packet injections from a host to the network.

Given these parameters, we define the non-blocking network packet storage S_{nnp} as the total number of packets that can be stored in the network without contention for the output ports of the switching nodes. It is given by adding the header, node routing and internode dela y times, multiplying this sum by R (counting the output data paths) and by N (the total number of SN nodes). Dividing the quantity by τ_P+1 gives

$$s_{nnp} = \frac{RN(\tau_H + \tau_D + \tau_L)}{(\tau_P + 1)} \ packets . \tag{1}$$

We now define the variable E_{ave} as the average number of hops needed to route a packet from a given source node to a given destination node. From [5], the average number of hops in the SN topology is given by

$$E_{ave} = \frac{3}{2}(k - 1) + 2^{1-k} \ hops \tag{2}$$

where k is the number of stages of switching nodes in the SN.

In addition to E_{ave}, one additional hop is needed to account for the time needed to route a packet through the node and link connected to the destination host. Given this condition, the total number of hops to route a packet from a source to a destination host is $(E_{ave}+1)$. Also, the packet duration, τ_P+1, represents the time needed by a packet at the switching node (connected to the destination host)

to finish routing through this node after setup of the data path within the node. The resultant total delay given by

$$\tau_{tot} = (E_{ave} + 1)(\tau_H + \tau_D + \tau_L) + (\tau_P + 1) \ time \ slots \tag{3}$$

and is the total time that a packet occupies any internal network resources, Specifically, this value is the average time from the insertion of the first bit of a packet into the switching node connected to the source host, to the evacuation of the last bit of that packet from the switching node connected to the destination host in the absence of contention. This value is different from the theoretical mean routing delay

$$\tau_{ave} = (E_{ave} + 1)(\tau_H + \tau_D + \tau_L) \ time \ slots \tag{4}$$

which is the average time needed for a packet to route from any source to any destination in the network in the absence of contention.

The next parameter we define is the average non-blocking network packet rate, C_{ave}. It specifies the maximum number of packets per time slot that can be sent through the network in a non-blocking fashion. It is given by

$$C_{ave} = \frac{S_{n\,np}}{(E_{ave} + 1)(\tau_H + \tau_D + \tau_L) + (\tau_P + 1)} \ packets/time \ slots \ . \tag{5}$$

This equation assumes that the traffic statistics are such that packets on average will require E_{ave} hops to reach their destination. If the traffic statistics are such that the majority of packets require the maximum number of hops to reach their destination, then Eq. (5) must be modified to use the maximum number of hops E_{max} as given by

$$E_{max} = 2k - 1. \tag{6}$$

For this case, we define the maximum non-blocking network packet rate, C_{max}, by

$$C_{max} = \frac{S_{nnp}}{(E_{max} + 1)(\tau_H + \tau_D + \tau_L) + (\tau_P + 1)} \text{ packets/time slots}, \quad (7)$$

The rate at which packets are input from a host to one physical port of a switching node is given by $1/(\tau_P+1)$. C_{ave} and C_{max} are governed by the routing delay through the network plus the evacuation time of packets from the destination node. The difference between these two rates is the reason why an imbalance exists between the input capacity and the internal network capacity for any recirculating topology with a single cylinder of node-to-node links. The SN topology minimizes E_{ave} and E_{max} (assuming no deflections) and is an ideal choice as a superset network architecture. MSN does not minimize the number of hops for a given number of nodes and this is the reason that the MSN is not as well suited as the SN to this type of scaling scheme.

For the SN topology, the expected number of hops grows linearly with the number of stages of switching nodes. By substituting Eq. (1) into Eq. (5), C_{ave} is given by

$$C_{ave} = \frac{RN(\tau_H + \tau_D + \tau_L)}{(\tau_P + 1)(E_{ave} + 1)(\tau_H + \tau_D + \tau_L) + (\tau_P + 1)^2} \text{ packets/time slots}. \quad (8)$$

The total rate at which packets are input to all N nodes from their respective hosts is given by

$$C_{in} = \frac{N}{\tau_P + 1} \text{ packets/time slots} \quad (9)$$

where N is just the total number of switching nodes in the SN.

To match the non–blocking network packet rate to the input packet rate C_{in}, we want C_{ave} to be equal to or larger than C_{in}. The ratio of C_{ave} to C_{in} is

$$\frac{C_{ave}}{C_{in}} = \frac{R}{(E_{ave} + 1) + (\tau_P + 1)/(\tau_H + \tau_D + \tau_L)} \qquad (10)$$

and a similar ratio can be written with C_{max} and C_{in}. Using the condition that these ratios must be greater than or equal to one, we find the minimum number of cylinders, R_{ave} (and R_{max}), needed for non–blocking operation is

$$R_{ave} \geq E_{ave} + 1 + (\tau_P + 1)/(\tau_H + \tau_D + \tau_L) \qquad (11)$$

and

$$R_{max} \geq E_{max} + 1 + (\tau_P + 1)/(\tau_H + \tau_D + \tau_L) . \qquad (12)$$

Equations (11) and (12) specify the minimum number of cylinders needed in terms of the average (maximum) number of hops (for a given topology), the packet duration τ_P, the header detection time τ_H, node setup and routing delay τ_D, and the node-to-node link delay τ_L. From Eq. (11), we see that R_{ave} is proportional to the number of expected hops for a given topology. This term is the dominant factor in Eq. (11) and is the motivation for desiring a topology which minimizes E_{ave} for a given number of nodes.

5 Network Simulations

We performed simulations of the MCSN architecture used as a WAN. The internal routing within the 3x3 nodes of Fig. 6 was implemented by Permutation Engines (PEs), which can perform non-blocking routing of the packet asynchronous data that would occur in this environment. The remainder of this section is organized as follows. Section 5.1 discusses the construction of the multi-cylin-

der SN model, Section 5.2 discusses several simulation model parameters. Section 5.3 presents some simulation results for the single and multi-cylinder SN architectures.

5.1 Simulation Model Definition

A 24-node MCSN was simulated using the SES/Workbench discrete event simulator [10]. The current model allows for an MCSN with PE switching nodes [9] for control and routing of packets through the SN. The model is composed of four layers: i) the traffic generator; ii) the protocol definition; iii) the network topology; and iv) the switching node model,

The traffic generator selects source and destination hosts at random with equal probability. The *source node* is the switching node connected to the source host; it receives new packets coming from the source into the network. The *destination node is* the switching node connected to the destination host; it removes packets which have finished routing from the network. Source nodes are selected from unused input links, while destination nodes are. selected without regard to existing traffic patterns.

The user can specify the size of a message in terms of the number of packets. All packets are of fixed length with the length being a user selectable parameter. The generator sends out the message, one packet at a time until the entire message has been sent. This mode of operation emulates an ATM style interface in that a message is broken down into smaller packets, each with its own header.

The three types of traffic patterns used – synchronous, skewed synchronous and asynchronous – are shown in Fig. 7. In this figure, the pulses represent the duration of one packet. For synchronous traffic, packets are injected into all switching nodes at the same time with a variable number of packet time slots allotted between such injections to achieve the desired network loading. For skewed synchronous traffic, single packets are injected into the SN at regular packet time slot intervals. Asynch-

ronous traffic is similar to skewed synchronous with a dither of several bit intervals about the mean injection points to simulate packet-asynchronous, bit-synchronous traffic.

The protocol definition layer specifies how packets are handled once a packet reaches an output port of a switching node. A deflection protocol makes three possible decisions: i) if the current node is the destination node and the host port is available, then route to the host; ii) if the current node is the destination node and the host port is not available, then recirculate to the current node; and iii) if the current node is not the destination node, then route to the next node, The first outcome results in sending a packet out of the network to the host. The second outcome will result when contention occurs at the destination node. Here, if the number of simultaneous packets which need to route to the host exceeds the number of output ports connected to the host, then one or more of these packets is recirculated back into the node. The final outcome will route a packet to the next desired node or deflect the packet to an undesired node.

The third layer of the model defines the interconnections between the switching nodes of the network. This layer defines the SN topology with the multi-cylinder augmentation. The topology layer provides the perfect shuffle interconnection between stages of switching nodes, and implements the routing algorithm (based on the packet header) for a single-cylinder SN as described in [5]. For the multi-cylinder SN, it also specifies the range of node output ports associated with the up and down output paths of a node in the single-cylinder SN.

The fourth layer of the model defines operation of the switching nodes of the network. This layer implements the PE control scheme for routing and contention resolution of asynchronous packets. The output of the topology layer determines which one of three types of output ports a packet desires. These outputs arc the top and bottom network outputs (which go to other nodes), and the outputs connected to the host as shown in Fig. 4. If two or more packets desire the same output port, one

packet will use the *primary* output port for the desired output port type and the other packet(s) will use one of the other output ports within the range for the output port type specified by the topology layer. For a network output port, a packet will continue to the next desired node but on a different cylinder of the network. For the host output port, the packet will either route to the host or be recirculated into the same node to try again.

The "Hot Potato" protocol in [5] is related to the protocol defined here and includes additional information about priority/age information in the packet header. This mechanism results in older packets having higher priority than newer packets when routing through a switching node. Packets in our SN model are not tagged with priority/age information for two reasons: i) header modification is difficult to do at gigabit/sec rates with optical signals; and ii) priority/age information is only useful for packet synchronous traffic. This added complication to the protocol was omitted because the switching nodes (and the network) are designed for packet asynchronous traffic.

5.2 Simulation Model Parameters

Both single and multi-cylinder SNS were simulated to verify the operation of the PE switching nodes, validate the SN simulation model, and assess the theoretical limitations of the multi-cylinder design approach. Tables 1 and 2 present analytical results from Eqs. (11) and (12) for the smallest integer value for the number of cylinders, R', needed for non-blocking operation of 24 and 2048 node MCSNS. Here R' corresponds to $\lceil R_{ave} \rceil$ or $\lceil R_{max} \rceil$ for E_{ave} and E_{max} respectively. We calculate this bound for packet lengths of 1 and 45 bits. The one-bit packet represents the minimum size packet and results in the fewest number of cylinders. These parameters are discussed in NO TAG for a single cylinder SN with age/priority information in the packet header. The 45-bit packet is a more practical case because it allows for 32 bits of header/data and 4-bit parity nibble with an 8-bit to 10-bit encoding scheme for fiber optic compatibility.

In Table 1, k=3 to give E_{ave}=3.25 and E_{max}=5 from Eqs. (2) and (6) respectively. In Table 2, k=8 to give E_{ave}=10.51 and E_{max}=15. In both tables, we assume τ_H and τ_L are equal to one bit time. The node delay τ_D for a pair of cascaded permutation engines (PE) with J physical input ports is $2J$ [9]. In this simulation we assumed $H+Q$ equal to R', so there area total of J=3R' input ports to the PE and anode delay of $6R'$ time slots. To solve for R_{ave} and R_{max} in Eqs. (11) and (12), we used a recursive procedure in which we initially assume that the term with τ_P is negligible and use the average and maximum number of hops to estimate R_{ave} and R_{max} respectively. Taking the next largest integer for R' results in a value for τ_D which can then be used to reevaluate $Rave$ and R_{max} *with* the packet length term. The results in Tables 1 and 2 were arrived at after one or two such iterations.

Tables 1 and 2 show analytically that longer length packets require more cylinders to provide non-blocking routing. Setting R' for the average number of hops guarantees that the internal network has the capacity to avoid blocking if packets require an average of E_{ave} hops. If all packets require& hops to reach their destination, more cylinders to accommodate the maximum number of hops must be used to avoid blocking. Table 2 shows a similar set of results for a 2048-node MCSN. The important point of comparison between these tables is the scaling of the routing delay through the MCSN with respect to the network input capacity. The average (and maximum) routing delay is directly proportional to k and scales as $\log_2 N$, where N is the number of nodes of the MCSN. From these tables, we see that the input capacity increases by a factor of 85 times, while the routing delay increases by only 5.7 times.

5.3 Simulation Results

Three different SN models were simulated to validate the MCSN concept, investigate the blocking characteristics of the single-cylinder network versus the MCSN architecture, and assess the theoretical limitations of the multi-cylinder design approach. The three models consist of a single-cylinder

64 node SN, and 5-cylinder and 6-cylinder 24 node MCSNs. All three models were simulated at various loading factors to determine the routing efficiencies for each configuration. Each simulation for a given set of parameters was run until approximately 1000 packets were correctly routed to their destinations. Figures 8 through 11 show the results of these simulations for synchronous and asynchronous packets. In these figures, routing delay in slots is plotted on the vertical axis and the average number of slots between injections is plotted on the horizontal axis, where an injection is defined as insertion of one or more packets into the network, Tables 3 through 5 show additional results.

In these figures, each light bar represents the mean routing delay and each dark bar the maximum routing delay for a given traffic load at the inputs of the network. In the simulation shown in Fig. 8, there were a total of 24 hosts and 64 nodes. Each host generates packets at its maximum possible rate for insertion into the network, thus only 24/64 or 37% of the total input capacity is utilized. In Fig. 8, the left most pair of bars depicts the mean and maximum delay, which increases without bound at this set of loading parameters. Similarly the left most pair of light and dark bars in Figs. 9 through 11 show the mean and maximum delay, for 24 hosts and 24 nodes, or when 100% of total input capacity is utilized.

The reason for defining a slot as the unit delay element is that these results can be scaled for different bit rates. If we assume a data rate of 1.2gigabit/sec, for example, the slot time is the reciprocal of this rate or 0.833 ns/slot. Hence in Figs. 10 and 11, for example, the maximum routing delay (for a traffic load of 100% of network input bandwidth) is 0.833 ns/slot multiplied by a delay of 342 slots for an actual delay of 285 ns.

For a 64 node SN, Eq. (2) gives a value of 4.62 for E_{ave} (the average number of hops a packet must make to reach its destination). The average number of node (plus link) delays for a packet to route correctly is $E_{ave}+1$ or 5.62 for this case. This parameter is converted to slots by using the expression

for the theoretical mean routing delay, $\tau_{ave}=(E_{ave}+1)(\tau_H+\tau_D+\tau_L)$, with $\tau_D=6$ for the single-cylinder SN and assuming $\tau_H=\tau_L=1$, Thus, the theoretical mean routing delay is 45 time slots.

For the single-cylinder 64-node SN with asynchronous packet traffic, Fig. 8 shows a mean delay of 51.5 slots with 32 slots pcr injection, which corresponds to 2% of network input capacity. For synchronous traffic at this loading factor, this model resulted in a mean routing delay of 129,1 slots. At 9% of input capacity with asynchronous traffic, this model exhibited a mean routing delay of 1001.9 slots or 22 times the theoretical mean, and 1345.1 slots or 30 times the theoretical mean for synchronous traffic. The maximum routing delay for asynchronous and synchronous traffic is 6395.5 and 7250 slots respectively, or 100 and113 times the theoretical maximum routing delay without deflection.

As can bc seen from these results and Fig, 8, the single--cylinder configuration is highly susceptible to packet deflection even at low network utilization. The routing delay when the number of slots per injection is small (meaning high network utilization), is directly related to the length of the simulation run. This situation is denoted by the arrows in Fig. 8 indicating that the routing delay for these cases is potentially unbounded.

The most striking result of the 64-node simulations is the large ratio of the maximum to mean routing delays even with light traffic loads. The reason for this situation is the blocking nature of the SN topology and the fact that no age/priority mechanism is being used to minimize the deflection probability as a packet circulates through the network. Since a packet deflection results in an additional k hops for a k–stage cyclic SN [6], a packet must avoid being deflected for at least three consecutive hops for a 24-node SN to reach its final destination. The end result is that scaling the network by increasing the number of nodes to avoid packet deflection is not optimal.

As can be seen in Fig. 8, a traffic load of 2% or more of network input capacity will increase the ratio of maximum to mean routing delay by two or more. Herein lies the desire to provide multiple SN cylinders. Since packet deflections will occur, by using multiple SN cylinders, even misrouted packets will still move closer to their destinations by using alternative paths in parallel to the desired path.

The results for the five-cylinder SN are shown in Fig. 9. At 100% of input capacity, the increase in the maximum routing delay for this model increased by a factor of two. The performance of the five-cylinder SN compared to the single-cylinder SN is significantly better duc to the additional cylinders of links. Note 4 in Fig. 9, however, shows that two packets were routed to the wrong destination host at 100% of input capacity. This result occurred because onl y five cylinders were used instead of the six required to match the input to network capacities.

By contrast the six–cylinder configuration results shown in Figs. 10 and 11, show only a 50% increase in the maximum routing delay through the network as the traffic load goes to 100% of the input capacity of the network. For 1.2 gigabit/sec data rates, the end result is an increase in the internal network routing delay from 0.19 μs to 0.28 μs. The six--cylinder configuration matches the input capacity of the network to the internal capacity of the network and avoids misrouting packets to the wrong destination. By setting, the number of cylinders equal to or greater than this ratio, the network has enough internal storage to hold the data in optical form until it reaches the destination node.

Tables 3, 4 and 5 below summarize these results with network input loading near 100% for the 24-node SNS. The results for the 64-node SNS correspond to loading at 20% and 70% of input capacity for packet sizes of onc and 45 slots respectively. The headings in these tables correspond to the number of MCSN nodes and cylinders, the packet duration in slots, and the theoretical and simulated mean and maximum routing delays in slots, where the theoretical maximum routing delay is $\tau_{max}=(E_{max}+1)(\tau_H+\tau_D+\tau_L)$. As can be seen from these tables, the network routing characteristics arc

similar for all three types of traffic patterns used. The additional point to note from these tables is the sensitivity of the single-cylinder SN to cell size. The maximum routing latencies shown for the single--cylinder 64-node SN in these tables with 45–bit long packets are limited by the simulation run time and arc potentially unbounded.

The results summarized in tables 3,4 and 5 assume $H=R'$ data paths to the host where $R'=\lceil R_{ave} \rceil$. Hence for the 6--cylinder 24--node SN, each switching node has six links to output data to its host. This particular model allows for up to six sources to send at 100% of the link rate to one destination simultaneously without contention.

Figure 12 shows the increase in the routing delay for the six- cylinder MCSN as the number of output data paths to a host is reduced from six to one. As can be seen in this figure, the routing delay increases by about a factor of five as the number of data paths drops to one. Reducing the number of data paths from six to three seems like a favorable tradeoff because the routing delay increases by only 50% while the amount of receiver hardware is reduced by 50%. As above, these simulation results used a uniform distribution to select unused inputs and any output (regardless of the current traffic load) while utilizing almost 100% of the input capacity of the network.

6 Conclusions

The results presented in section 5.3 show that the multi-cylinder network can handle synchronous, skewed synchronous and asynchronous types of traffic. For a single–cylinder SN, the maximum routing delay can be very large compared to the mean delay even at low network loading. This result is due to the high deflection probability of the single-cylinder SN. The MCSN, on the other hand, can run at 100% input capacity with minimal routing delay due to input to network capacity matching afforded by adding extra cylinders with dynamic interaction between all cylinders, This fully interconnected multi-cylinder configuration is the essence of the MCSN.

The above results also show that the MCSN has routing characteristics that are relatively unaffected by the timing of traffic entering the network. This network property is very important for an all-optical data path WAN due to the difficulty in synchronizing optical packets as they enter the network from geographically distributed sources.

The parameters which affect network performance include packet duration, expected number of hops, node setup and routing delay, and node-to-node link delays. By proper selection of these parameters, traffic congestion in the network can be alleviated using extra routing cylinders to avoid packet deflections. Furthermore, imbalances in the traffic pattern can be alleviated by using bandwidth augmentation to allow multiple sources to send simultaneous packets to the same destination. From Fig. 12, wc sce that using two links to transfer data from a switching node to its host nearly halves the network routing delay compared to using only one link.

By using multiple routing cylinders and bandwidth augmentation, this architecture can handle traffic loads up to 100% of network input capacity. If, however, fewer routing cylinders and/or output ports arc desired (to reduce system cost), source throttling can be used to avoid congestion problems within the network and at the output ports of the network. The MCSN has many design options to obtain the desired performance as a function of network complexity. In the future, we will carry out additional simulations to further characterize these relationships.

7 Acknowledgements

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the Ballistic Missile Defense Organization, Innovative Science and Technology Office, and the National Aeronautics and Space Administration. The work was performed as part of JPL's Center for Space Microelectronics Technology.

8 Technical References

[1] N. F. Maxemchuk, "Comparison of Deflection and Store-and-Forward Techniques in the Manhattan Street and Shuffle-Exchange Networks," *IEEE INFOCOM '89,* pp. 800-809, April 1989.

[2] A. K. Gupta and N. D. Georganas, "Analysis of a Packet Switch with Input and Output Buffers and Speed Constraints," *IEEE INFOCOM '91,* pp. *694-700, 1991.*

[3] M. Sidi and W. Z. Liu, "Congestion Control Through Input Rate Regulation," *IEEE INFOCOM '89,* pp. *17641768, 1989.*

[4] I. Rubin and K. D. Lin, "Input Rate Flow Control for High-Speed Networks: Source vs. Switch Level Performance Trade-off," *IEEE GLOBECOM '91*, pp. 249-253, 1991.

[5] J. R, Sauer, "An Optoelectronic Multi–Gb/s Packet Switching Network," *OCS Technical Report 89–06*, February 1989.

[6] A. Krishna and B. Hajek, "Performance of Shuffle–Like Switching Networks with Deflection," *IEEE INFOCOM '90,* pp. *473-480, 1990.*

[7] 1.. N. Bhuyan, "Interconnect ion Networks for Parallel and Distributed Processing", *IEEE Comp.,* vol. 20110.6, pp. 9-12, 1987.

[8] G. F. Lcv, N. Pippenger and L. G. Valiant, "A Fast Parallel Algorithm for Routing in Permutation Networks", *IEEE Trans. on Comp., vol.* C--30, pp. 93--100, 1981.

[9] S. P. Monacos and A. A. Sawchuk, "A Permutation Engine Switching Node", submitted to the J. of Parallel and Distributed Computing.

[10] SES/workbench User's Manual, Release 2.0, Scientific and Engineering Software Inc., Austin, Texas, 1991.

9 List of Tables

1 .

Packet length (time slots) τ_P	Average no. of hops E_{ave}	Maximum no. of hops E_{max}	Node routing delay (time slots) τ_D	No. of SN cylinders R'
1	3.25	–	30	5
1	–	5	42	7
45	3.25	–	36	6
45	–	5	48	8

Table 1

Packet length (time slots) τ_P	Average no. of hops E_{ave}	Maximum no. of hops E_{max}	Node routing delay (time slots) τ_D	No. of SN cylinders R'
1	10.51	15	72	12
1	10.51	15	102	17
45	10.51	15	78	13
45	10.51	15	108	18

Table 2

No. of nodes	Cylinders R'	Pkt size	Mean delay τ_{ave}	Max delay τ_{max}	Mean delay	Max delay
			Theoretical		Simulation	
24^{1}	5	1	136	" 2	130.6	288
$64^{2,4}$	1	1	45	64	140.0	690
24^{1}	6	45	161.5	228	159.4	342
$64^{3,4}$	1	45	45	64	207.2	1497

Notes: 1)These results are at max input capacity for 24-node SN.
2) This case is at 9% of input capacity for a 64-node SN.
3) This case is at 7096 of input capacity for a 64- node SN.
4) These cases have traffic. misrouted to wrong destination.

Table 3 Asynchronous traffic comparison.

No. of nodes	Cylinders R'	Pkt size	Mean delay τ_{ave}	Max delay τ_{max}	Mean delay	Max delay
			Theoretical		Simulation	
24^{1}	5	1	136	192	134.5	192
$64^{2,4}$	1	1	**45**	**64**	131.4	719.6
24^{1}	6	45	**161.5**	**228**	162.6	342
$64^{3,4}$	1	45	45	64	174.7	1855

Notes: I)These. results are at rnax input capacity for 24–node SN.
2) This case is at 9% of input capacity for a 64--node SN.
3) This case is at 70% of input capacity for a 64–node. SN.
4) These cases have traffic misrouted to wrong destination.

Table 4 Synchronous skewed traffic comparison.

No. of nodes	Cylinders R'	Pkt size	Mean delay τ_{ave}	Max delay τ_{max}	Mean delay	Max delay
			Theoretical		Simulation	
24^{1}	5	1	136	192	135.3	224
$64^{2,4}$	1	1	45	64	137.1	721.7
24^{1}	6	45	161,5	228	162,4	342
$64^{3,4}$	1	45	45	64	283.5	1686

Notes:)) These results are at max input capacity for 24- node SN.
2) This case is at 9% of input capacity for a 64 –node SN.
3) This case is at 70% of input capacity for a 64– node SN.
4) These. cases have traffic misrouted to wrong destination.

Table 5 Synchronous traffic comparison.

List of Figures

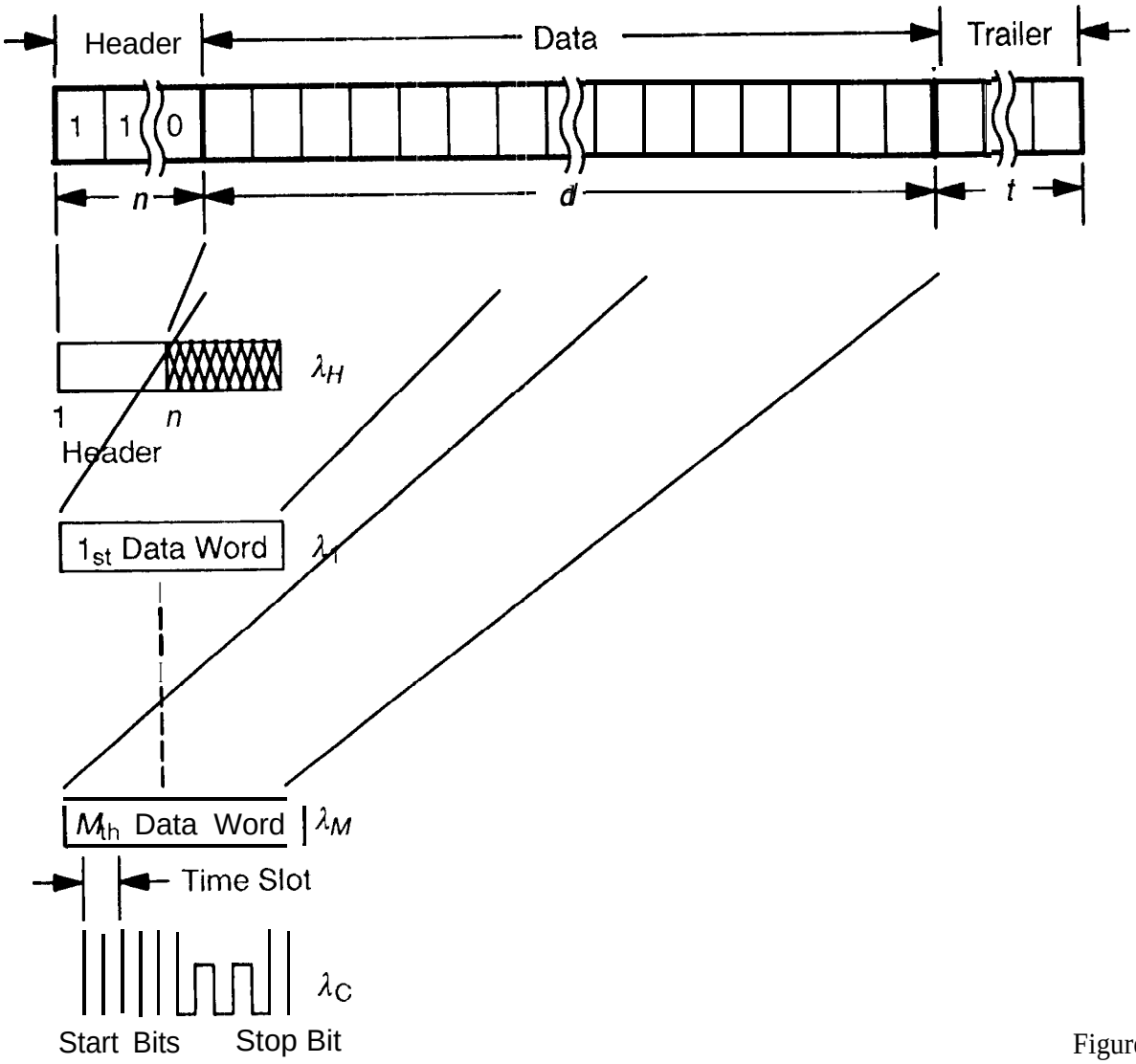

Header
Data
Trailer
1
1
0
n
d
t
λ_H
1
n
Header
1st Data Word
λ_1
Mth Data Word
λ_M
Time Slot
λ_C
Start Bits
Stop Bit

Figure 1

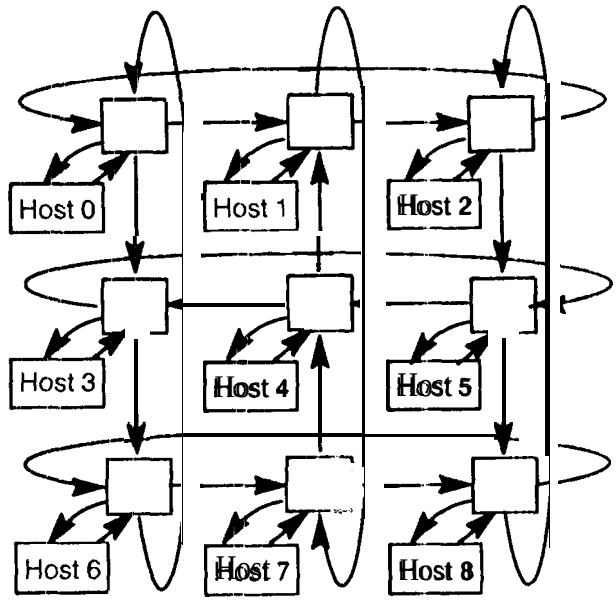


Figure 2

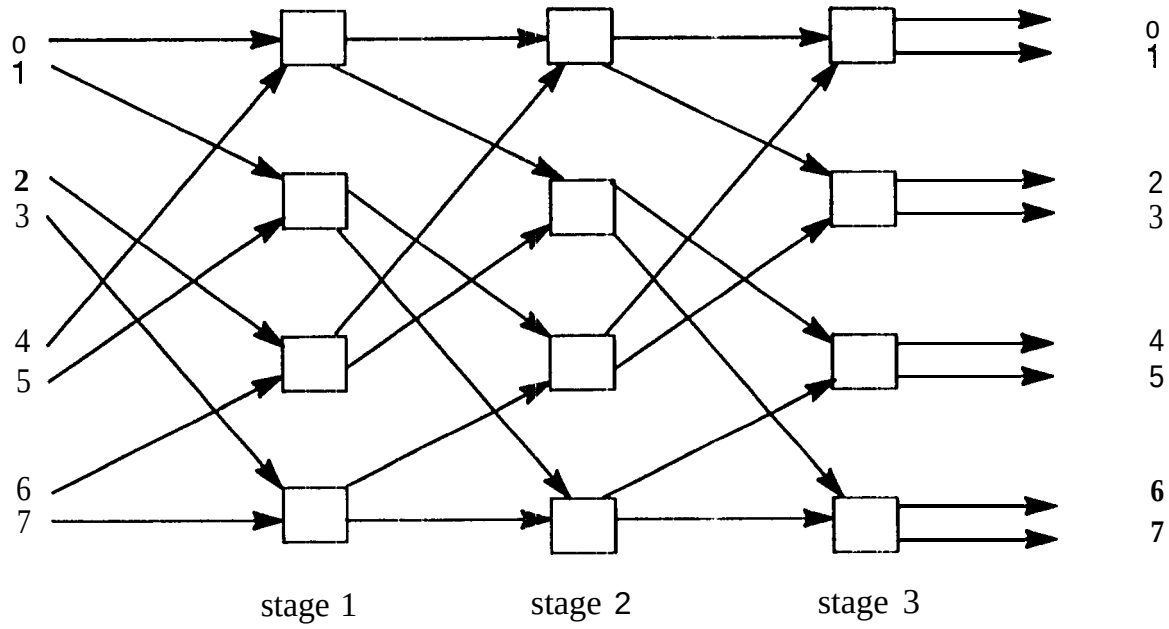


Figure 3

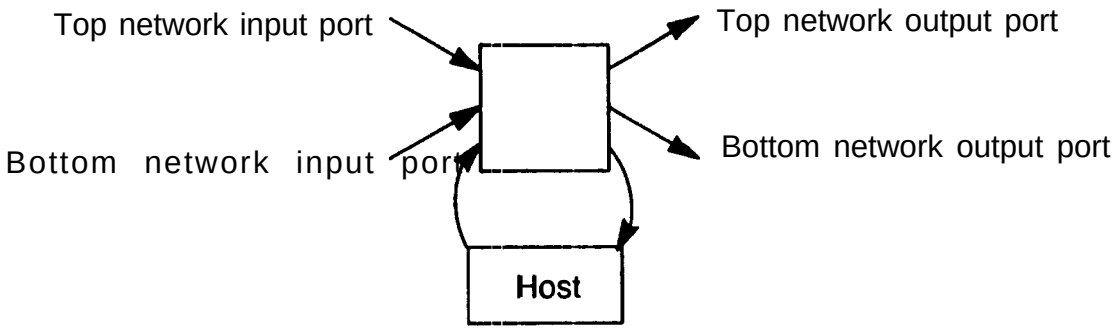

Top network input port
Top network output port
Bottom network input port
Bottom network output port
Host

Figure 4

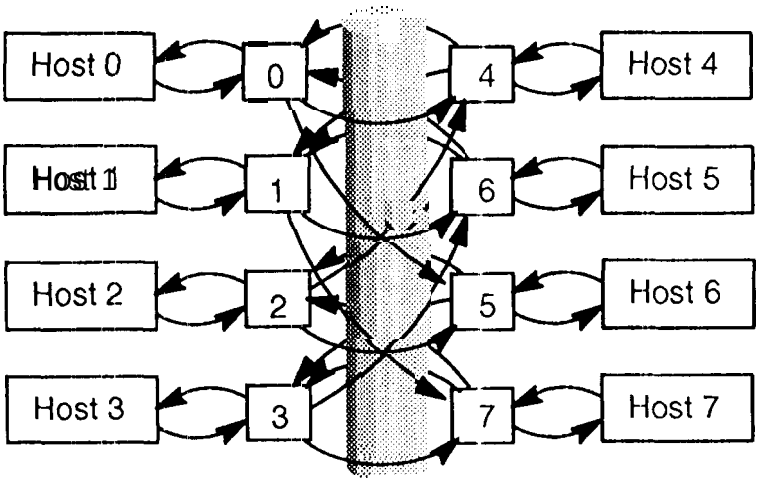


Figure 5

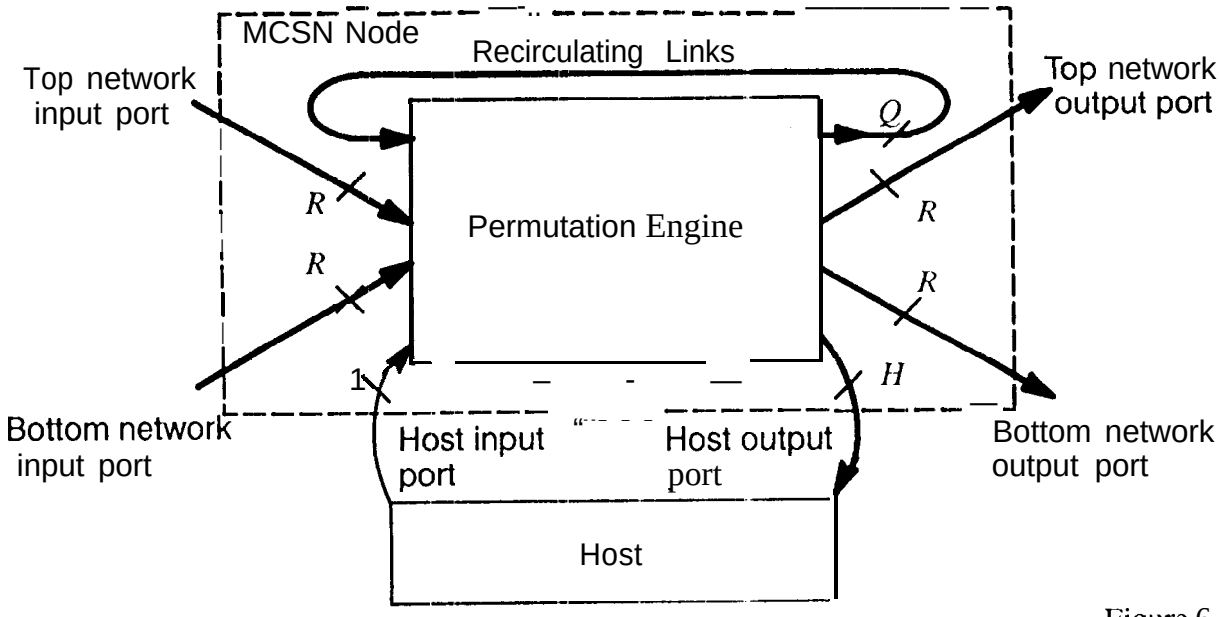

MCSN Node
Recirculating Links
Top network input port
Top network output port
Q
R
R
R
R
Permutation Engine
1
H
Bottom network input port
Host input port
Host output port
Bottom network output port
Host

Figure 6

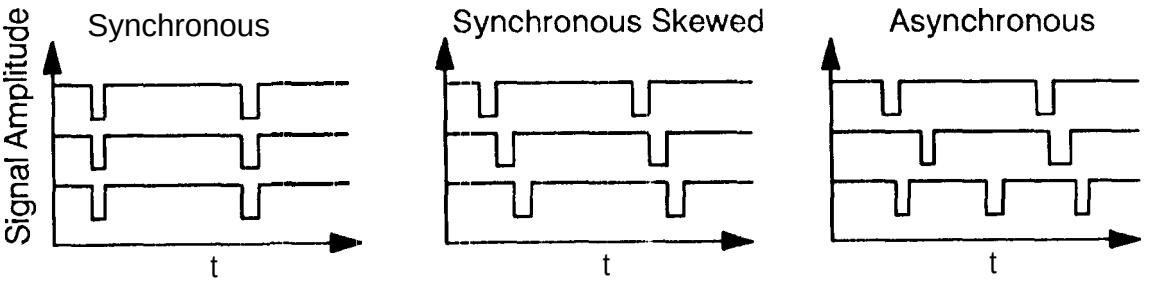

Signal Amplitude
Synchronous
t
Synchronous Skewed
t
Asynchronous
t

Figure 7

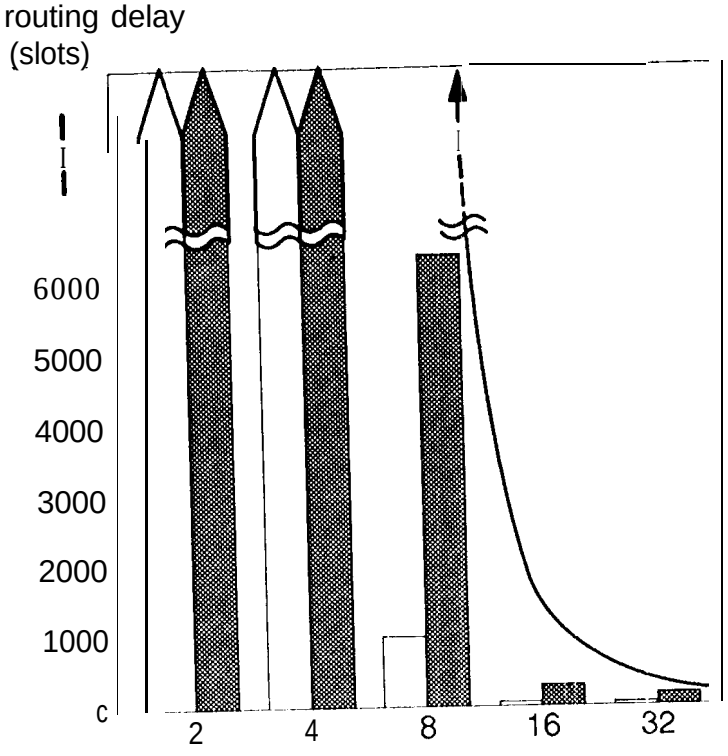
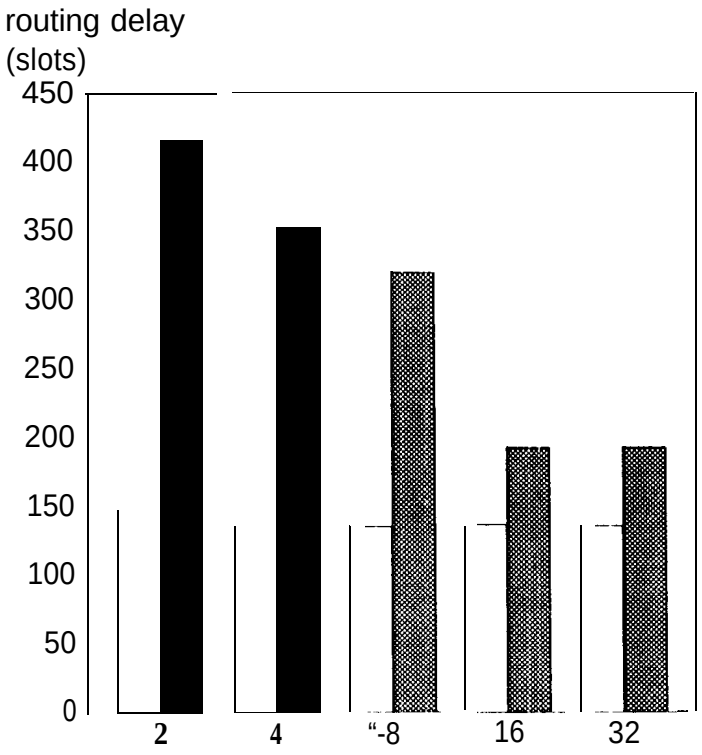

routing delay
(slots)
6000
5000
4000
3000
2000
1000
0
2
4
8
16
32
B mean
max
Notes:
1. Packet asynchronous traffic
2. Packet duration = 45 slots
3. One packet per injection
4. Many misroutes occurred
average number of slots between injections

Fig. 8



Notes:

1, Packet asynchronous traffic

2. Packet duration = 45 slots

3. One packet per injection

4. Two misroutes occurred

Fig. 9

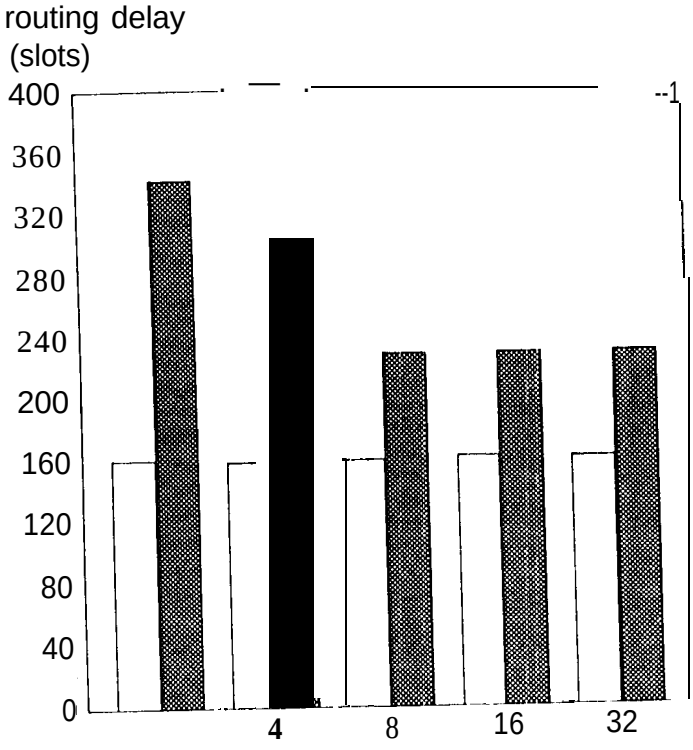
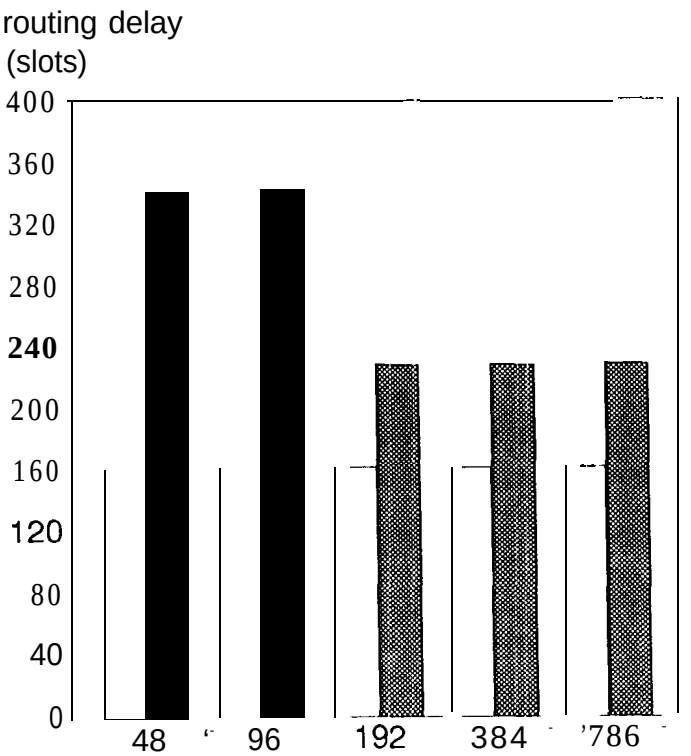

routing delay
(slots)
400
360
320
280
240
200
160
120
80
40
0
4
8
16
32
mean
max
Notes:
1. Packet asynchronous traffic
2. Packet duration = 45 slots
3. One packet per injection
4. No misroutes occurred
average number of slots between injections

Fig. 10



Notes:

1. Packet asynchronous traffic
2. Packet duration = 45 slots
3. 3.24 packets per injection
4. No misroutes occurred

Fig, 11

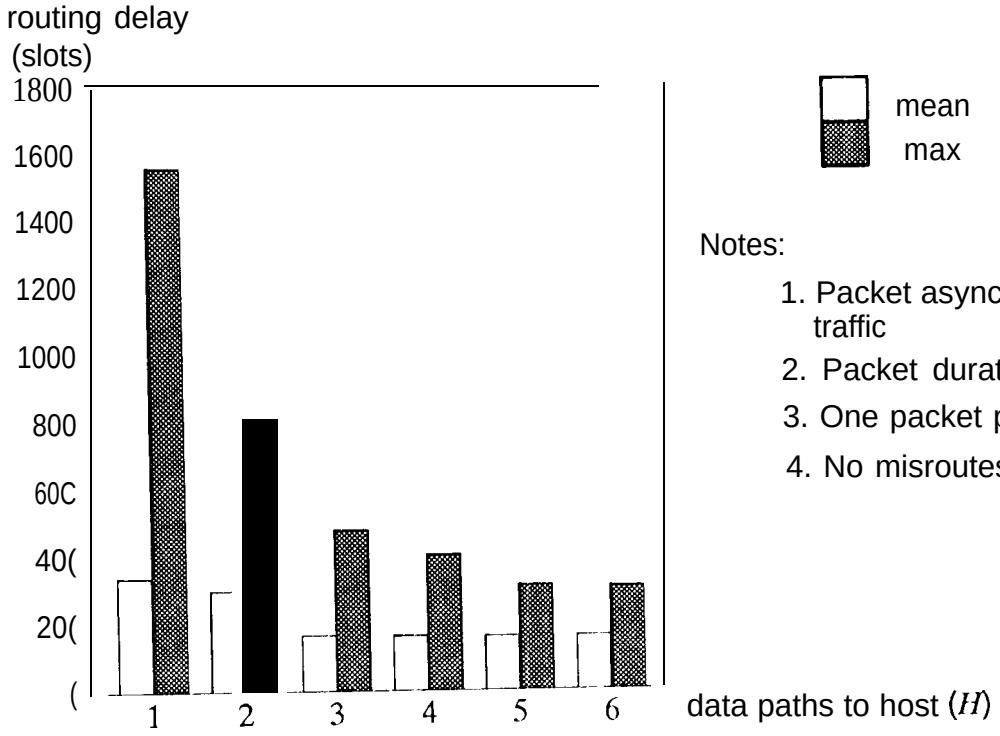

routing delay
(slots)
1800
1600
1400
1200
1000
800
60C
40(
20(
(
1
2
3
4
5
6
data paths to host (H)
mean
max
Notes:
1. Packet asynchronous traffic
2. Packet duration =45 slots
3. One packet per injection
4. No misroutes occurred

Fig. **12**